Unsupervised attack pattern detection in cyber-security using Bayesian topic modelling

Francesco Sanna Passino[1†], Anastasia Mantziou[1,2], Philip Thiede[1], Ross Bevington[3], Nick Heard[1]
[1]Imperial College London – [2]The Alan Turing Institute, London – [3]MSTIC, Microsoft Corporation
[†]f.sannapassino@imperial.ac.uk

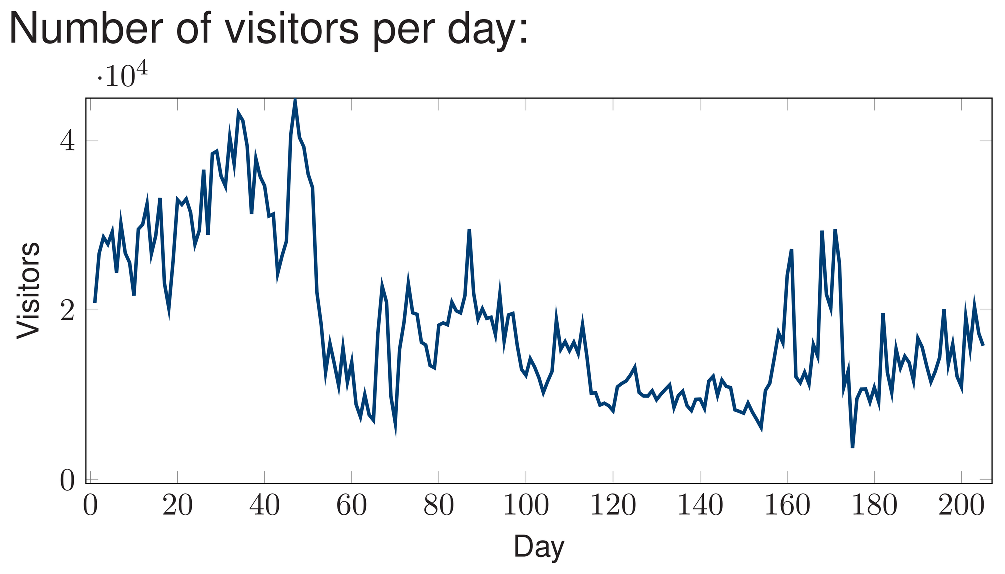
1. Motivation: modelling honeypots

- **Statistical cyber-security** is still in relative infancy.
- Most research to date has been concerned with **anomaly and intrusion detection**:
  - Build statistical models of **normal behaviour** of some aspects of an enterprise network;
  - Leverage the cyber-defender's advantage: **intimate knowledge of their own network**;
  - Requires limited knowledge of an attacker's intention, implying some robustness to different attacks;
  - Significance tests could lack power if they do not match the **current threat landscape**.
- A honeypot is a **decoy system** designed to be attacked and lure attackers into revealing themselves.
- Microsoft crafts **legitimate-looking honeypot systems** to avoid detection and extract maximal information.
  - "Clean room" bash/Linux simulator;
  - Every password is correct (eventually)!
  - Support a variety of protocols;
  - Injecting faults to tease out more interactions, moving the attacker outside their preferred path.
- Microsoft monitors their network of honeypots to identify **emerging threats** from thousands of daily attacks.
- Information for each **session**:
  - Time of connection;
  - IP address;
  - OS and window size;
  - Credentials;
  - Clipboard contents;
  - Protocol and port;
  - **Commands**.

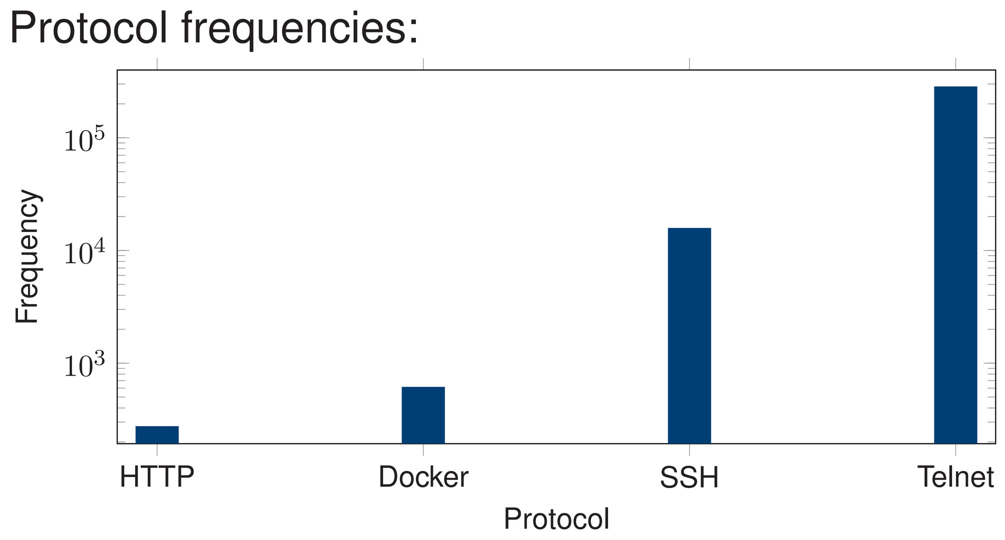
- Example of a session:

```
1628687161.212426 1628687162.525791 209.141.54.197 Tcp SSH
↪ 22 US ?? ['cd /tmp || cd /var/run || cd /mnt || cd
↪ /root || cd /; wget http://107.175.94.7/wget.sh; curl
↪ -O http://107.175.94.7/wget.sh; chmod 777 wget.sh; sh
↪ wget.sh; tftp 107.175.94.7 -c get tftp1.sh; chmod 777
↪ tftp1.sh; sh tftp1.sh; tftp -r tftp2.sh -g
↪ 107.175.94.7; chmod 777 tftp2.sh; sh tftp2.sh; ftpget
↪ -v -u anonymous -p anonymous -P 21 107.175.94.7 ftp.sh
↪ ftp.sh; sh ftp.sh; rm -rf wget.sh tftp1.sh tftp2.sh
↪ ftp.sh; rm -rf *']
```

- Number of visitors per day:



- Protocol frequencies:



2. Objective: clustering sessions

- We would like to **cluster** the honeypot sessions according to the attackers' **intentions**. This is an **unsupervised learning problem**, with an unknown number of classes.
- Appealing to analogies in text analysis, **latent Dirichlet allocation models** provide a natural framework.
- In general, there are **three main difficulties**:
  1. **Tokenisation of commands** into *words*, dealing with analogies for ***stop-words*** and ***misspellings***.
     - Regular expressions, splitting on $/;|-;$
     - Wildcarding exotic URLs and HEX sequences.
  2. Topic models are **unidentifiable** and inference is plagued by **convergence difficulties**.
  3. Topic models typically assume that all documents are **non-zero mixtures** of a **fixed** number of topics. Ideally we want **each overarching topic** to correspond to **one hacking group or behaviour**.

Acknowledgements

This work is funded by the **Microsoft Security AI** research grant *"Understanding the enterprise: Host-based event prediction for automatic defence in cyber-security"*.

3. Proposed methodology: clustering via Bayesian topic modelling

- Suppose we observe D documents (sessions) and define:
  - N_d – number of commands in session d;
  - $M_{d,j}$ – number of words in command j of session d;
  - $w_{d,j,i} \in V$ – ith word in the the jth command of document d;
  - V – observed vocabulary.
- Let $\boldsymbol{\xi}_{d,j,i} \in \mathbb{R}^{|V|}$ denote the probability mass function of $w_{d,j,i}$ over V, such that:

$$w_{d,j,i} \sim \boldsymbol{\xi}_{d,j,i}.$$

- A range of **topic model structures** for $\boldsymbol{\xi}_{d,j,i}$ is considered. Two examples are:
  1. **Hierarchical**: Each session topic is a distribution on command-level topics $\implies$ **two layers of latent topics**.
  2. **Constrained**: Each session has a primary topic and a **global secondary** topic.
- Let $\boldsymbol{t} = (t_1, \dots, t_D)$ where $t_d \in \{1, \dots, K_{\max}\}$ denotes the index of the overarching topic of session d, and $K_{\max}$ is a hypothetical maximum number of topics (this can, for example, be set equal to the number of documents).
- Let $\boldsymbol{\lambda} \in \mathbb{R}^{K_{\max}}$ be a probability mass function on the topic indices $\{1, \dots, K_{\max}\}$, so

$$t_d \sim \boldsymbol{\lambda}, \quad d = 1, \dots, D.$$

- The topics $\boldsymbol{t}$ are the object of inferential interest $\Rightarrow$ **latent attacker's intent**.

4. Hierarchical topic models

- **Two layers** of topics:
  1. **Command** topic indices, $s_{d,j}$. Each command topic $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{H_{\max}}$ is a distribution over V.
  2. **Document** topic indices, t_d. Each document topic $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{K_{\max}}$ is a distribution over command topics.
- Let $\boldsymbol{\Psi}$ be the $H_{\max} \times |V|$ matrix with j-th row $\boldsymbol{\psi}_j$, and $\boldsymbol{\Phi}$ the $K_{\max} \times H_{\max}$ matrix with k-th row $\boldsymbol{\xi}_k$. Then, marginally:

$$\boldsymbol{\xi}_{d,j,i} = \boldsymbol{\lambda}^\intercal \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\Psi}.$$

- More specifically,

$$\begin{aligned} \boldsymbol{\psi}_k &\sim \text{Dirichlet}(\boldsymbol{\zeta}),\ k = 1, \dots, K_{\max}, \\ \boldsymbol{\phi}_h &\sim \text{Dirichlet}(\boldsymbol{\eta}),\ h = 1, \dots, H_{\max}, \\ s_{d,j} \mid t_d, \{\boldsymbol{\psi}_k\} &\sim \boldsymbol{\psi}_{t_d}, \\ w_{d,j,i} \mid s_{d,j}, \{\boldsymbol{\phi}_h\} &\sim \boldsymbol{\phi}_{s_{d,j}}, \end{aligned}$$

where $i = 1, \dots, M_{d,j},\ j = 1, \dots, N_d,\ d = 1, \dots, D$.

5. Primary and secondary topics

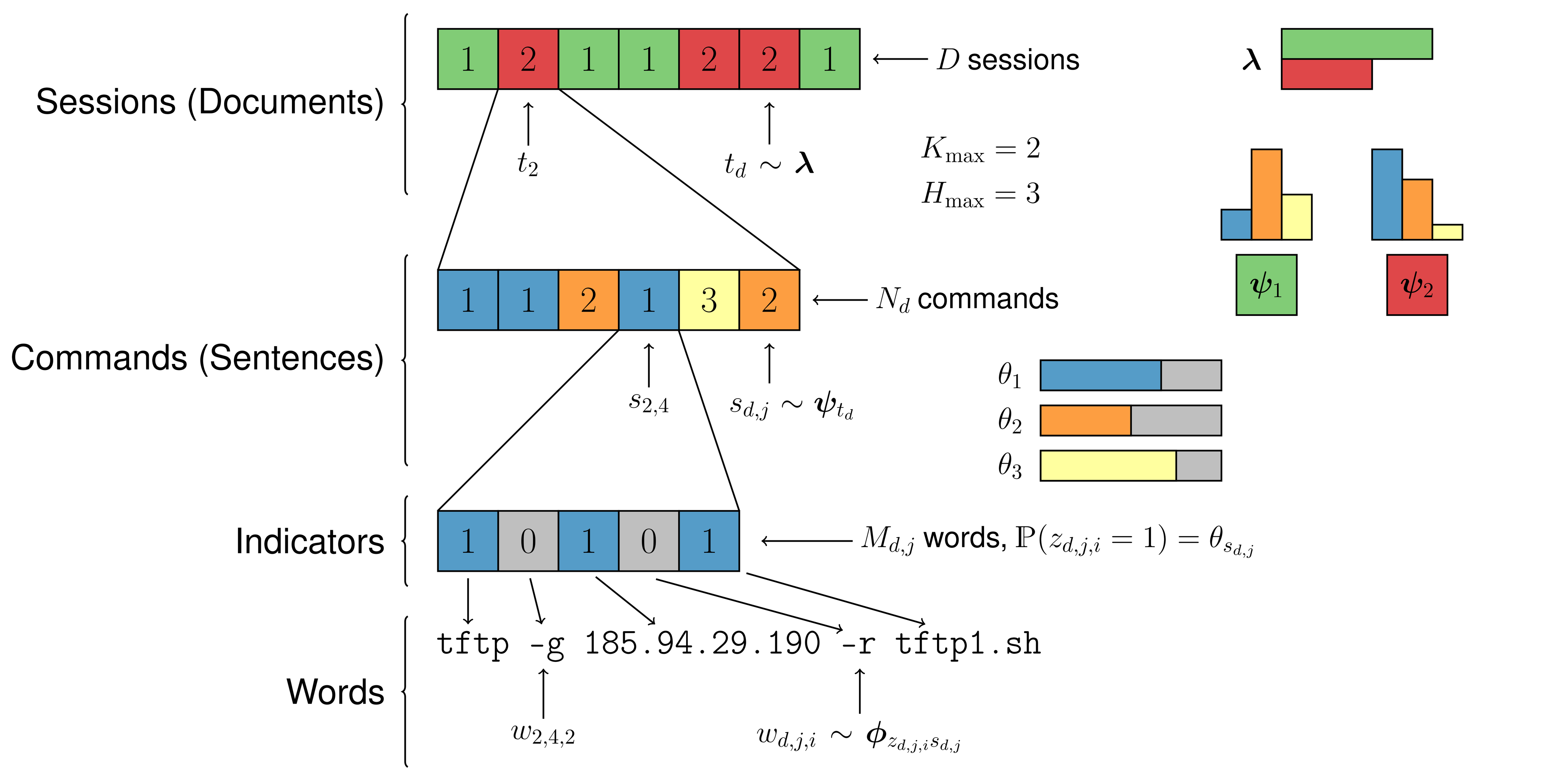
- A global topic 0 forms a **baseline topic** shared by all documents as their secondary topic. This could represent **uninteresting, navigational commands**.
- A Bernoulli indicator variable $z_{d,j,i}$ determines whether each word is drawn from the **primary document topic** or the **background secondary topic**.
- More specifically,

$$\begin{aligned} \boldsymbol{\phi}_k &\sim \text{Dirichlet}(\boldsymbol{\eta}), \\ \theta_k &\sim \text{Beta}(\alpha_k, \alpha_0), \\ z_{d,j,i} \mid \theta_d &\sim \text{Bernoulli}(\theta_d), \\ w_{d,j,i} \mid z_{d,j,i}, t_d, \{\boldsymbol{\phi}_k\} &\sim \boldsymbol{\phi}_{t_d z_{d,j,i}}, \end{aligned}$$

where $i = 1, \dots, M_{d,j},\ j = 1, \dots, N_d,\ d = 1, \dots, D$ and $k = 0, 1, 2, \dots, K_{\max}$.
- The two approaches can also be **combined**:

$$w_{d,j,i} \mid z_{d,j,i}, s_{d,j}, \{\boldsymbol{\phi}_h\} \sim \boldsymbol{\phi}_{z_{d,j,i} s_{d,j}}.$$

6. Schematic combination of hierarchical and constrained topic models



7. Results

- Inference is performed via **collapsed Metropolis-within-Gibbs sampling**, with **split-merge** moves.
- Promising results, with some meaningful clusters.
- Uncovered a **previously undocumented bot searching for coin miners**, then published on the MS Security blog.

Cluster	Content
1	MIRAI, Mozi
2	MIRAI
3	(ptmx) unnamed botnet, SBIDIOT
4	MIRAI
5	MIRAI, (ptmx) unnamed botnet
6	Bushido
7	MIRAI, (ptmx) unnamed botnet
8	MIRAI, Shellbot
9	Mikrotik bot
10	***Interesting***
11	MIRAI, SDITIOT
12	MIRAI, (ptmx) unnamed botnet
13	Coin miners, (ptmx) unnamed botnet, Hive attacking bot
14	Mikrotik bot
15	Coin mining, IP scanning, General recon

8. Discussion

- Honeypot data are currently an **underused** data source.
- **Unsupervised classification** of sessions is challenging.
- Goals are to find:
  - A compact representation which aids **identifiability**;
  - An accompanying inference algorithm which addresses **convergence** issues.
- All models discussed this poster assume a fixed size $|V|$ of the vocabulary, and a fixed number of session-level and command-level topics, $K_{\max}$ and $H_{\max}$ respectively.
  - Problematic if the model is used for clustering future sessions $\Rightarrow$ an ***infinite* vocabulary** should be used.
  - *New* attack patterns or intents arise $\Rightarrow$ **unbounded** number of session-level and command-level topics.

$$\boldsymbol{\lambda} \sim \text{GEM}(\gamma), \quad \boldsymbol{\psi}_k \sim \text{GEM}(\tau), \quad \boldsymbol{\phi}_\ell \sim \text{GEM}(\eta).$$

- Upcoming paper presents all models and BNP extensions in more details, with simulations and results on real-data.
- ***python* library** available at `fraspass/lda_clust`.